BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37230
doi: 10.3897/biss.3.37230

OPEN ACCESS

Conference Abstract

Training Machines to Identify Species using GBIF-mediated Datasets

Tim Robertson[‡], Serge Belongie[§], Hartwig Adam[|], Christine Kaeser-Chen[¶], Chenyang Zhang[¶], Kiat Chuan Tan[¶], Yulong Liu[¶], Denis Brulé[#], Cédric Deltheil[#], Scott Loarie[¤], Grant Van Horn[¤], Oisin Mac Aodha[«], Sara Beery[«], Pietro Perona[«], Kyle Copas[‡], John Thomas Waller[‡]

‡ Global Biodiversity Information Facility, Copenhagen, Denmark
§ Cornell Tech, New York, United States of America
| Google Research, Los Angeles, United States of America
¶ Google Research, New York, United States of America
# Google Research, Paris, France
¤ iNaturalist, San Francisco, United States of America
« California Institute of Technology, Pasadena, United States of America

Corresponding author: Tim Robertson (trobertson@gbif.org)

Received: 12 Jun 2019 | Published: 19 Jun 2019

Citation: Robertson T, Belongie S, Adam H, Kaeser-Chen C, Zhang C, Chuan Tan K, Liu Y, Brulé D, Deltheil C, Loarie S, Van Horn G, Mac Aodha O, Beery S, Perona P, Copas K, Waller J (2019) Training Machines to Identify Species using GBIF-mediated Datasets. Biodiversity Information Science and Standards 3: e37230.
https://doi.org/10.3897/biss.3.37230

Abstract

Advances in machine vision technology are rapidly enabling new and innovative uses within the field of biodiversity. Computers are now able to use images to identify tens of thousands of species across a wide range of taxonomic groups in real time, notably demonstrated by iNaturalist.org, which suggests species IDs to users (https://www.inaturalist.org/pages/computer_vision_demo) as they create observation records. Soon it will be commonplace to detect species in video feeds or use the camera in a mobile device to search for species-related content on the Internet.

The Global Biodiversity Information Facility (GBIF) has an important role to play in advancing and improving this technology, whether in terms of data, collaboration across teams, or citation practice. But in the short term, the most important role may relate to initiating a cultural shift in accepted practices for the use of GBIF-mediated data for training of artificial intelligence (AI).

© Robertson T et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

“Training datasets” play a critical role in achieving species recognition capability in any machine vision system. These datasets compile representative images containing the explicit, verifiable identifications of the species they include. High-powered computers run algorithms on these training datasets, analysing the imagery and building complex models that characterize defining features for each species or taxonomic group. Researchers can, in turn, apply the resulting models to new images, determining what species or group they likely contain. Current research in machine vision is exploring (a) the use of location and date information to further improve model results, (b) identification methods beyond species-level into attribute, character, trait, or part-level ID, with an eye toward human interpretability, and (c) expertise modeling for improved determination of “research grade” images and metadata.

The GBIF community has amassed one of the largest datasets of labelled species images available on the internet: more than 33 million species occurrence records in GBIF.org have one or more images (https://www.gbif.org/occurrence/gallery). Machine vision models, when integrated into the data collection tools in use across the GBIF network, can improve the user experience. For example, in citizen science applications like iNaturalist, automated species suggestion helps even novice users contribute occurrence records to GBIF.

Perhaps most importantly, GBIF has implemented uniform (and open) data licensing, established guidelines on citation and provided consistent methods for tracking data use through the Digital Object Identifiers (DOI) citation chain. GBIF would like to build on the lessons learned in these activities while striving to assist with this technology research and increase its power and availability.

We envisage an approach as follows:

1. To assist in developing and refining machine vision models, GBIF plans to provide training datasets, taking effort to ensure license and citation practice are respected. The training datasets will be issued with a DOI, and the contributing datasets will be linked through the DOI citation graph.

2. To assist application developers, Google and Visipedia plan to build and publish openly-licensed models and tutorials for how to adapt them for localized use.

3. Together we will strive to ensure that data is being used responsibly and transparently, to close the gap between machine vision scientists, application developers, and users and to share taxonomic trees capturing the taxon rank to which machine vision models can identify with confidence based on an image’s visual characteristics.

Keywords

biodiversity species, machine vision, training models

Presenting author

Tim Robertson

Presented at

Biodiversity_Next 2019